Genomics 97 (2011) 379–385

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Methods

Generalized *F* accelerated failure time model for mapping survival trait loci

Xiaojing Zhou [a], Li Yan [b], Daniel R. Prows [c], Runqing Yang [d,e,*]

[a] Department of Mathematics, Heilongjiang Bayi Agricultural University, Daqing 163319, People's Republic of China
[b] College of Information Technology, Heilongjiang Bayi Agricultural University, Daqing 163319, People's Republic of China
[c] Division of Human Genetics, Cincinnati Children's Hospital Medical Center and University of Cincinnati College of Medicine, Cincinnati, OH 45229, USA
[d] College of Animal Science and Veterinary Medicine, Heilongjiang Bayi Agricultural University, Daqing 163319, People's Republic of China
[e] School of Agriculture and Biology, Shanghai Jiaotong University, Shanghai 200240, People's Republic of China

ARTICLE INFO

Article history:
Received 20 October 2010
Accepted 3 February 2011
Available online 26 February 2011

Keywords:
Survival trait
Accelerated failure time model
Generalized *F* distribution
Interval mapping

ABSTRACT

As the two most popular models in survival analysis, the accelerated failure time (AFT) model can more easily fit survival data than the Cox proportional hazards model (PHM). In this study, we develop a general parametric AFT model for identifying survival trait loci, in which the flexible generalized *F* distribution, including many commonly used distributions as special cases, is specified as the baseline survival distribution. EM algorithm for maximum likelihood estimation of model parameters is given. Simulations are conducted to validate the flexibility and the utility of the proposed mapping procedure. In analyzing survival time following hyperoxic acute lung injury (HALI) of mice in an F_2 mating population, the generalized *F* distribution performed best among the six competing survival distributions and detected four QTLs controlling differential HALI survival.

1. Introduction

Lander and Botstein [1] have first proposed the interval mapping method for simultaneously estimating the positions and effects of quantitative trait loci (QTL), and since then, various extensions have been developed [2–7]. Earlier statistical methods for mapping QTL mainly focused on the continuous and normally distributed traits. With the need to reveal genetic mechanisms for extensive complex diseases and economic traits in plants, animals and humans, researchers must exploit additional approaches to map QTL for the traits with skewed distributions, such as binary, ordinal, count and survival traits.

As time-to-event, survival traits are broadly defined as the length of time between two events. These traits with long right tails do not follow the normal distribution and are often subject to censoring, therefore the existing statistical methods for mapping QTL have difficulties analyzing such traits appropriately. Some methodologies of survival analysis, including the cure-rate model, parametric and semi-parametric models, are developed sequentially. In the parametric and semi-parametric models, the Cox proportional hazard model (PHM) or the accelerated failure time model (AFT) is the natural choice for genetic association of survival time with markers and linkage analysis for mapping survival time loci [8].

Symons et al. [9] formulated the QTL effects on the failure time with a PHM and estimated the model parameters and computed LOD scores by a variant of the EM algorithm [10]. Diao [11] also developed a PHM with a Weibull baseline hazard function to characterize the effects of QTL on the failure time. In the case of such a spike in the phenotype distribution, Broman [12] used a two-part parametric model and a nonparametric approach based on the Kruskal–Wallis test for QTL mapping. Fine et al. [13] proposed nonparametric estimates for genetic effects of QTL, which complemented the rank tests of Kruglyak and Lander [14]. Based on the PHM, Diao and Lin [15] proposed efficient likelihood-based inference measures and developed semi-parametric statistical methods for mapping survival trait loci. Using simulated data with different structures, Moreno et al. [16] systematically compared the parametric model based on Weibull distribution, semi-parametric model and classical interval mapping based on the normal distribution. Fang [17] investigated a simple and efficient approach to estimating QTL parameters through partial likelihood function. In outbred populations, the variance component model based on methods of Epstein et al. [18] or Pankratz et al. [19] are appropriate for mapping QTL of survival traits.

In survival analysis, the AFT model has an intuitive physical interpretation for real-life examples as it directly expresses the failure time rather than the probability as in the PHM and therefore would be an important alternative to the PHM [20,21]. The AFT model makes modeling simple as it relates the logarithm of the failure time linearly to the covariates [22,23]. It also reduces the potential error amplifications from linking models with different structures. In contrast to PHM, the AFT model is rarely used in QTL mapping. Cheng and Tzeng [8] proposed parametric and semi-parametric methods based on AFT models for interval mapping but the fact that using the likelihood derived by Diao et al. [11] to estimate model

* Corresponding author at: College of Animal Science and Veterinary Medicine, Heilongjiang Bayi Agricultural University, Daqing 163319, PR China. Fax: +86 459 6819206.
E-mail address: runqingyang@sjtu.edu.cn (R. Yang).

0888-7543/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2011.02.003

parameters greatly increases the computational burden. Meanwhile, extensive simulations have revealed that the parametric estimators may be more efficient in determining the effect and location of QTL, although parametric estimators may have obvious bias when selecting the incorrect error distribution. In contrast, the semi-parametric inference is robust to the error distribution. There is no apparent difference in statistical power of QTL detection between parametric and semi-parametric estimations [8].

Similar to the PHM, the AFT model describes the relationship between survival probabilities and a set of covariates. For each error distribution in AFT model, there is a corresponding survival distribution [24]. Many error distributions in AFT model are available, such as commonly used exponential distribution, Weibull distribution, log-normal distribution, gamma distribution and so on. However, these are just special forms of the generalized *F* distribution [23]. The objectives of this study are (i) to formulate a general parametric model for mapping survival trait loci based on AFT model with generalized *F* distribution; (ii) to give the EM algorithm of maximum likelihood estimation for QTL parameters; and (iii) to demonstrate the flexibility and statistical power of the proposed method by using simulations and applying the method to identify QTL for survival time following hyperoxic acute lung injury (HALI) in an F_2 mating population of mice.

2. Methods

2.1. Genetic model

We describe our method in the context of an F_2 population, in which there are three genotypes QQ, Qq, and qq. However, it can be easily extended to other kinds of crosses (e.g., backcross, recombination inbred lines, and four-way crosses). All n individuals were observed for survival time (T: t_1,t_2,…,t_n) and genotyped for markers with a known linkage map. Assume that a single QTL flanked by any two adjacent markers M_k and M_{k+1} is responsible for the traits of interest and specify that the QTL multiplicatively act on the failure time T or, additively, on log$T$, then the AFT model for mapping survival trait loci can be described by

$$y_i = \mu + z_i a + w_i d + \sigma\varepsilon_i \tag{1}$$

where $y_i = \log T_i$ for the ith individual, μ is population mean, and a and d are the additive and dominance effects of QTL, respectively. z_i and w_i are genotype indicator variables related to genetic effects a and d, respectively, which are defined as:

$$z_i = \begin{cases} 1 \\ 0 \\ -1 \end{cases} \text{ and } w_i = \begin{cases} 0 & \text{for QQ} \\ 1 & \text{for Qq} \\ 0 & \text{for qq} \end{cases}$$

σ is scale parameter and ε_i is a random error which is assumed to follow the generalized *F* distribution.

2.2. Generalized F distribution

If the random error ε has an *F* distribution with $2\nu_1$ and $2\nu_2$ degrees of freedom, the density function of generalized *F* is then

$$f(\varepsilon) = (\nu_1 e^{\varepsilon}/\nu_2)^{\nu_1}(1 + \nu_1 e^{\varepsilon}/\nu_2)^{-(\nu_1 + \nu_2)} B(\nu_1, \nu_2)^{-1} \tag{2}$$

and the survival function is

$$S(\varepsilon) = \int_0^{\nu_2(\nu_2 + \nu_1 e^{\varepsilon})^{-1}} x^{\nu_2 - 1}(1-x)^{\nu_1 - 1} B(\nu_2, \nu_1)^{-1} dx \tag{3}$$

where $\nu_1 > 0$, $\nu_2 > 0$ and $B(\nu_1, \nu_2)$ is the beta function evaluated at ν_1 and ν_2.

Table 1
Relationship between generalized *F* and other commonly used distributions.

4 parameters	3 parameters	2 parameters	1 parameter
Generalized *F*	*F* ($\sigma=1$)	Gamma ($\nu_2\to\infty$)	χ^2 ($\nu_1 e^{-\mu}=0.5$) Exponential ($\nu_1=1$)
	Extended Generalized Gamma ($\nu_1\to\infty$ *or* $\nu_2\to\infty$)	Gamma ($\sigma=1$)	χ^2 ($\nu_1 e^{-\mu}=0.5$) Exponential ($\sigma=1$)
		Weibull ($\nu_1=1$ $\nu_2\to\infty$)	Exponential ($\sigma=1$) Rayleigh ($\sigma=0.5$)
		Inverse Weibull ($\nu_2=1$, $\nu_1\to\infty$)	
		Log normal ($\nu_1=1$, $\nu_2\to\infty$)	
	Generalized Log-logistic ($\nu_1=\nu_2=\nu$)	Log-logistic ($\nu=1$)	
	Burr III ($\nu_2=1$)	Log-logistic ($\nu_1=1$)	
	Burr XII ($\nu_1=1$)	Log-logistic ($\nu_2=1$)	

As shown in Hogg and Ciamp [25] and Kalbfleisch and Prentice [23], the generalized *F* distribution includes many commonly used distributions as special cases, such as the Weibull, log-normal, gamma, and log-logistic distributions, and so on. Table 1 displays some specific values of generalized *F* distribution parameters and associated distribution.

2.3. Maximum likelihood estimation

The survival function of T can be expressed by the survival function of ε_i:

$$S(t_i) = S_{\varepsilon_i}\left(\frac{\log t_i - \mu - z_i a - w_i d}{\sigma}\right) \tag{4}$$

According to the relationship between survival function and density function $f(t) = -S'(t)$, we can obtain the density function for survival time as:

$$f(t_i) = \frac{1}{\sigma t_i} f_{\varepsilon_i}\left(\frac{\log t_i - \mu - z_i a - w_i d}{\sigma}\right) \tag{5}$$

where f_{ε_i} is the density function of the random variable ε_i. Assume that there are some right-censoring records in the observed survival times, let C_i be censoring time for the ith individual and $\Delta_i = I(T_i \le C_i)$, where $\Delta_i = 1$ when T_i is fully observed (uncensored), otherwise, $\Delta_i = 0$. For the ith individual, then, the survival density function can be formulated as

$$\varphi(t_i) = f(t_i)^{\Delta i}\ S(t_i)^{1-\Delta i} \tag{6}$$

All the survival density functions for the uncensored record and the records with different censoring types are given in Discussion. Given the three QTL genotypes, density function $\varphi(t_i)$ will have three types, denoted as $\varphi(t_{i|l})$, where $l = 1, 2, 3$ corresponds to QTL genotypes QQ, Qq, and qq, respectively. Let $p_{i|l}$ be the conditional probabilities of the above three QTL genotypes given the flanking markers M_k and M_{k+1}, then a mixture model [26] can be formed as

$$\varphi(t_i) = \sum_{l=1}^{3} p_{i|l}\varphi\left(t_{i|l}\right). \tag{7}$$

Suppose that the trait values are independent of each other, a likelihood corresponding to data t_1,t_2,…,t_n is the product of

independent mixture models on n individuals, given the survival time (T) and marker information (M), that is

$$L_1(\beta|T,M) = \prod_{i=1}^{n} \varphi(t_i) = \prod_{i=1}^{n} \left(\sum_{l=1}^{3} p_{i|l} \varphi\left(t_{i|l}\right) \right) \tag{8}$$

where the vector $\beta = (\mu,a,d,\sigma,\theta,\delta)^T$ with θ being the parameters in baseline hazard function and δ being the scanning position.

Then, the log-likelihood is given by

$$L(\beta|T,M) = \sum_{i=1}^{n} log\left(\sum_{l=1}^{3} p_{i|l} \varphi\left(t_{i|l}\right) \right) \tag{9}$$

with derivatives

$$\begin{aligned} \frac{\partial L(\beta|T,M)}{\partial \beta} &= \sum_{i=1}^{n} \sum_{l=1}^{3} \frac{p_{i|l} \varphi\left(t_{i|l}\right)}{\sum_{l=1}^{3} p_{i|l} \varphi\left(t_{i|l}\right)} \frac{\partial}{\partial \beta} log\, \varphi\left(t_{i|l}\right) \\ &= \sum_{i=1}^{n} \sum_{l=1}^{3} p^{*}_{j|l} \frac{\partial}{\partial \beta} log \varphi\left(t_{i|l}\right), \end{aligned} \tag{10}$$

where we define $p^{*}_{j|l} = \frac{p_{i|l} \varphi\left(t_{i|l}\right)}{\sum_{l=1}^{3} p_{i|l} \varphi\left(t_{i|l}\right)}$, $l = 1, 2, 3$, as the posterior probabilities of three QTL genotypes for jth individual. Herein, the EM algorithm [27] is implemented to solve the maximum likelihood estimations of β.

The iteration steps are described below:

(1) Initialize values $\beta^{(0)} = (\mu^{(0)},a^{(0)},d^{(0)},\sigma^{(0)},\theta^{(0)},\delta)^T$ for $\beta = (\mu,a,d,\sigma,\theta,\delta)^T$.
(2) Compute the posterior probabilities $p^{*}_{j|l}$ ($l = 1, 2, 3$) given the initial values.
(3) Solve for $\frac{\partial}{\partial \beta} \log L\,(\beta|T,M) = 0$ to get the estimates of β, denoted as $\beta^{(1)} = (\mu^{(1)},a^{(1)},d^{(1)},\sigma^{(1)},\theta^{(1)},\delta)^T$. In practical computation, the simplex algorithm implemented with function 'fminsearch' in the MatLab can be used to obtain the solution for the nonlinear and complicated equations.
(4) Replace the initial parameters $\beta^{(0)} = (\mu^{(0)},a^{(0)},d^{(0)},\sigma^{(0)},\theta^{(0)},\delta)^T$ by $\beta^{(1)} = (\mu^{(1)},a^{(1)},d^{(1)},\sigma^{(1)},\theta^{(1)},\delta)^T$ and go back to step (2).
(5) Iterate until a criterion of convergence is reached. At the convergence, the values of the parameters are the maximum likelihood (ML) solutions, denoted by $\hat{\beta} = (\hat{\mu},\hat{a},\hat{d},\hat{\sigma},\hat{\theta},\delta)^T$.

2.4. Significance test

A likelihood ratio statistic was used to test the significance of the QTL effect. Substituting the above maximum likelihood estimates to equation (9), we first obtained the log-likelihood value under the full model as L_1 ($\hat{\mu},\hat{a},\hat{d},\hat{\sigma},\hat{\theta},\delta|T,M$) and then evaluated the log-likelihood function under the null model (reduced model) so that $a = d = 0$ was used in place of a and d, denoted by L_0 ($\hat{\mu}_0,\hat{\sigma}_0,\hat{\theta}_0,\delta|T,M$). Note that $\hat{\mu}_0$, $\hat{\sigma}_0$, and $\hat{\theta}_0$ are different from $\hat{\mu}_0$, $\hat{\sigma}_0$, and $\hat{\theta}_0$ because the former are estimated by maximizing

$$\begin{aligned} logL(\mu,\sigma,\theta|T,M) = &\sum_{i=1}^{n} \Delta_i\, log\left[\frac{1}{\sigma t_i} f_{\varepsilon_i}\left(\frac{logt_i - \mu}{\sigma} \right) \right] \\ &+ \sum_{i=1}^{n} (1-\Delta_i)\, log\left[S_{\varepsilon_i}\left(\frac{logt_{iR} - \mu}{\sigma} \right) \right], \end{aligned}$$

the log-likelihood function under the reduced model.

The likelihood ratio test statistic is defined as

$$\mathrm{LOD} = -2\log_{10} \frac{L_0\left(\hat{\mu}_0,\ \hat{\sigma}_0,\ \hat{\theta}_0,\delta|T,M\right)}{L_1\left(\hat{\mu},\ \hat{a},\ \hat{d},\ \hat{\sigma},\ \hat{\theta},\delta|T,M\right)}$$

To determine the significance of the LOD test, we used permutation tests to evaluate the critical threshold [28]. First, a number of permuted samples were generated by repeatedly shuffling the relationships between marker genotypes and phenotypes. Then, a series of the maximum LODs was calculated for each sample. Finally, the critical threshold was obtained from the distribution of the maximum LODs.

By the same way, we calculated the LOD statistic at each locus over the genome (by spacing of 1 or 2 cM) and plotted the profile for LODs against the linkage map distance. The linkage map position corresponding to a peak of the LOD plot was determined as the maximum-likelihood estimate of the QTL location.

3. Simulations

To investigate the operating characteristics of the proposed methods in practical situations, we performed simulation studies using an F_2 design. A chromosome with a total length of 100 centiMorgans (cM) was considered, on which eleven equally spaced co-dominant markers were simulated with sample size of 150 and 300. A single QTL was put at position 25 cM between markers 3 and 4. The genetic effect of the QTL was designated at two levels of a (0.10 and 0.15) without d. A scaled parameter σ was taken to be 0.5 and the interval-mapping step size was set at 1 cM.

Survival times were generated from the model (1) based on the Weibull distribution, which is described in detail as follows: given density function of the random variable ε in Table 2, its distribution function is derived as $1 - \exp(-\exp(\varepsilon))$. According to the method by Mood et al. [29], random numbers of ε can be generated from $\varepsilon = \ln(-\ln(1-U))$, where $\ln(-\ln(1-U))$ is the inverted function of ε's distribution function and U is a random variable following a uniform distribution. Substituting ε into model (1), we then simulated survival time with $\exp(\mu + za + wd + \sigma\varepsilon)$ under the given scenario, in which records of 15% were censored randomly.

The experiment was replicated 100 times to estimate QTL parameters and to access the statistical power of QTL detection under each survival distribution and scenario simulated. Critical values of the test statistic used to declare statistical significance differed due to the survival distribution used in the mapping model. These significance thresholds were determined by simulating 500 additional samples under the null model with $a = 0$ and $d = 0$, based

Table 2
The commonly used error distributions and corresponding survival distributions.

Error		Survival time	
Error distribution	Density function	Survival distribution	Survival function
Extreme value (1 parameter)	$\exp(\varepsilon - \exp(\varepsilon))$ $(\sigma = 1)$	Exponential	$\exp(-\lambda t)$
Extreme value (2 parameters)	$\exp(\varepsilon - \exp(\varepsilon))$	Weibull	$\exp[-(\lambda t)^{\gamma}]$
Normal	$\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-\varepsilon^2}{2}\right)$	Log-normal	$1 - \Phi(\gamma \log \lambda t)$
Log-gamma	$\frac{\exp(k\varepsilon - \exp(\varepsilon))}{\Gamma(k)}$ $(\sigma = 1)$	Gamma	$1 - \frac{\int_0^{\lambda t} x^{k-1} e^{-x} dx}{\Gamma(k)}$
Logistic	$\frac{\exp(\varepsilon)}{(1 + \exp(\varepsilon))^2}$	Log-logistic	$\frac{1}{1 + (\lambda t)^{\gamma}}$

on the generalized *F* distribution. Statistical power of QTL detection was calculated as the percentage of the number of those simulations in which a significant QTL was detected.

We analyzed the simulated data by using mapping models based on Weibull and generalized *F* distributions, respectively. Parameter estimates (standard deviations) and the statistical power obtained with the two mapping models are listed in Table 3. With either Weibull or generalized *F* distributions, there is little bias between estimates and true values for QTL effects and location. Because the two mapping models had the same statistical power to detect QTL, this implied that the generalized *F* distribution does not over-fit the data set generated from the Weibull distribution. The estimated precision of parameters and statistical power of QTL detection, as expected, increased with the QTL effect, and sample size increased for each survival distribution. Like the Weibull distribution, the generalized *F* distribution can also fit the simulated data with four other survival distributions (results not shown), demonstrating the flexibility of the generalized *F* distribution in mapping survival trait loci with the parametric AFT model.

4. Example

A mouse model system for mapping QTL of hyperoxic acute lung injury (HALI) survival has been established by crossing sensitive B (C57BL/6J) strain mice and significantly more resistant S (129X1/SvJ) strain mice to HALI mortality[30,31]. The reciprocal F_1 lines were first generated by mating B females to S males (BS) and S females to B males (SB). The 4 possible F_2 crosses were systematically bred through BS×BS, BS×SB, SB×BS, and SB×SB (female F_1 listed first) intercross mating schemes. A total of 840 F_2 mice were phenotyped for survival time in hours and genotyped for 97 polymorphic microsatellite markers distributed over the genome, including the X chromosome. Total segregation ratio for the genotyped markers was 1.023:1.935:1.000 and Chi-square test showed that segregation ratios for each marker accorded with the expected value of 1:2:1. The logarithms of raw survival times were adjusted for the effects of each system environment factor due to dam, sire, and sex.

We analyzed the data using the AFT models with the generalized *F* distribution and the exponential, Weibull, log-normal, gamma, and log-logistic distributions. The survival distributions and the corresponding error distributions are summarized in Table 2. We used the Bayesian information criterion (BIC) [32] as the model selection criterion of the best error distribution function. The BIC is defined as

$$\mathrm{BIC} = -2\log L\left(\hat{\beta}\,|T,M\right) + \mathrm{dimension}(\beta|T,M)\ log(n),$$

where $\hat{\beta}$ is the maximum likelihood estimation of β under the reduced model, $(\beta|T,M)$ represents the number of independent parameters under this model and n is the number of observations. The best error distribution function is the one that displays the minimum BIC value, performing the characteristic of maximum likelihood and parsimonious parameters. Since these commonly used distributions are special cases of the generalized *F* distribution, the likelihood ratio test can be conducted based on nested models for model selection. The corresponding statistic logarithm of likelihood ratio is denoted as

$$log-\mathrm{LR} = -2[\ logL\left(\hat{\beta}\,|T,M,\mathrm{generalized}\right) - logL\left(\hat{\beta}\,|T,M,\mathrm{commonly\ used}\right)],$$

which follows a Chi-square distribution with the *df* degree of freedom, where *df* equals the difference in number of parameters between the compared distributions.

Table 4 tabulates log-likelihood, the BIC, and the log-LR values for each survival distribution. It can be seen that different survival distributions give different BIC values. In general, the generalized *F* distribution performs better than the five commonly used distributions, although BIC values for the log-logistic distribution and gamma distribution are close to that of the generalized *F* distribution. With likelihood ratio test, we further found that the goodness of fit to the data set by the generalized *F* distribution was significantly higher than that by the other survival distributions. Here, the critical threshold of Chi-square distribution with the maximum freedom degree of 3 was 7.92 at the significance level of 5%.

Figure 1 plots the profiles of LOD test statistics over the genome under the six competing survival distributions. The genome-wide empirical critical thresholds for significance declaration are obtained by using permutation tests with 1000 replicates, which are 3.8, 5.6, 7.2, 2.6, 5.0 and 3.5 for generalized *F*, exponential, Weibull, log-normal, Gamma and log-logistic distributions, respectively, at the 5% significance level. With the generalized *F* distribution, four significant QTLs were identified on chromosomes 1, 4 and 15. Parameter estimates of the QTLs with generalized *F* function are listed in Table 5. In the F_2 population, with these parameter estimates, we can draw three survival curves corresponding to QTL genotypes (Fig. 2). Comparing the three curves for each QTL, all four detected QTLs led to the change of survival density, where the difference in the shape of survival density function among the three QTL genotypes were similar for the third and fourth QTLs. As compared to the other survival distributions, mapping analysis based on the generalized *F* distribution can identify more QTLs and, in this case,

Table 3
Parameter estimates (standard deviations) and statistical powers obtained with interval mapping based on Weibull and generalized *F* distributions for the simulated data from Weibull distribution.

Sample size	True effect	$a=0.10$		$a=0.15$	
	Distribution	Weibull	Generalized *F*	Weibull	Generalized *F*
300	σ	0.50 (0.02)	0.44 (0.11)	0.49 (0.02)	0.47 (0.08)
	μ	2.00 (0.04)	2.01 (0.05)	1.99 (0.03)	2.00 (0.03)
	α	0.11 (0.02)	0.11 (0.02)	0.15 (0.03)	0.15 (0.02)
	d	0.02 (0.01)	0.02 (0.02)	0.01 (0.01)	0.01 (0.01)
	ν_1		0.879 (0.32)		0.94 (0.29)
	ν_2		5.87×10^5 (4.49×10^5)		7.17×10^5 (5.13×10^5)
	Position	28.86 (7.74)	27.89 (6.58)	25.63 (4.38)	25.68 (4.12)
	Power	79%	79%	99%	99%
500	σ	0.50 (0.02)	0.46 (0.07)	0.50 (0.02)	0.48 (0.04)
	μ	1.99 (0.03)	2.01 (0.04)	2.00 (0.03)	2.00 (0.03)
	α	0.10 (0.02)	0.11 (0.02)	0.15 (0.02)	0.15 (0.02)
	d	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
	ν_1		0.95 (0.30)		0.97 (0.21)
	ν_2		6.98×10^5 (4.12×10^5)		7.47×10^5 (3.65×10^5)

Table 4
The log-likelihood, the BIC and the log-LR values under different survival distributions for hyperoxic acute lung injury survival time in mice.

Distribution	Log*L*	Number of parameters	BIC	Log-LR
Generalized *F*	−118.7	4	264.3	
Exponential	−874.3	1	1755.3	1511.2
Weibull	−236.2	2	485.9	235.0
Log-normal	−594.6	2	1202.7	951.8
Gamma	−145.1	2	303.7	52.8
Log-logistic	−126.3	2	266.1	15.2

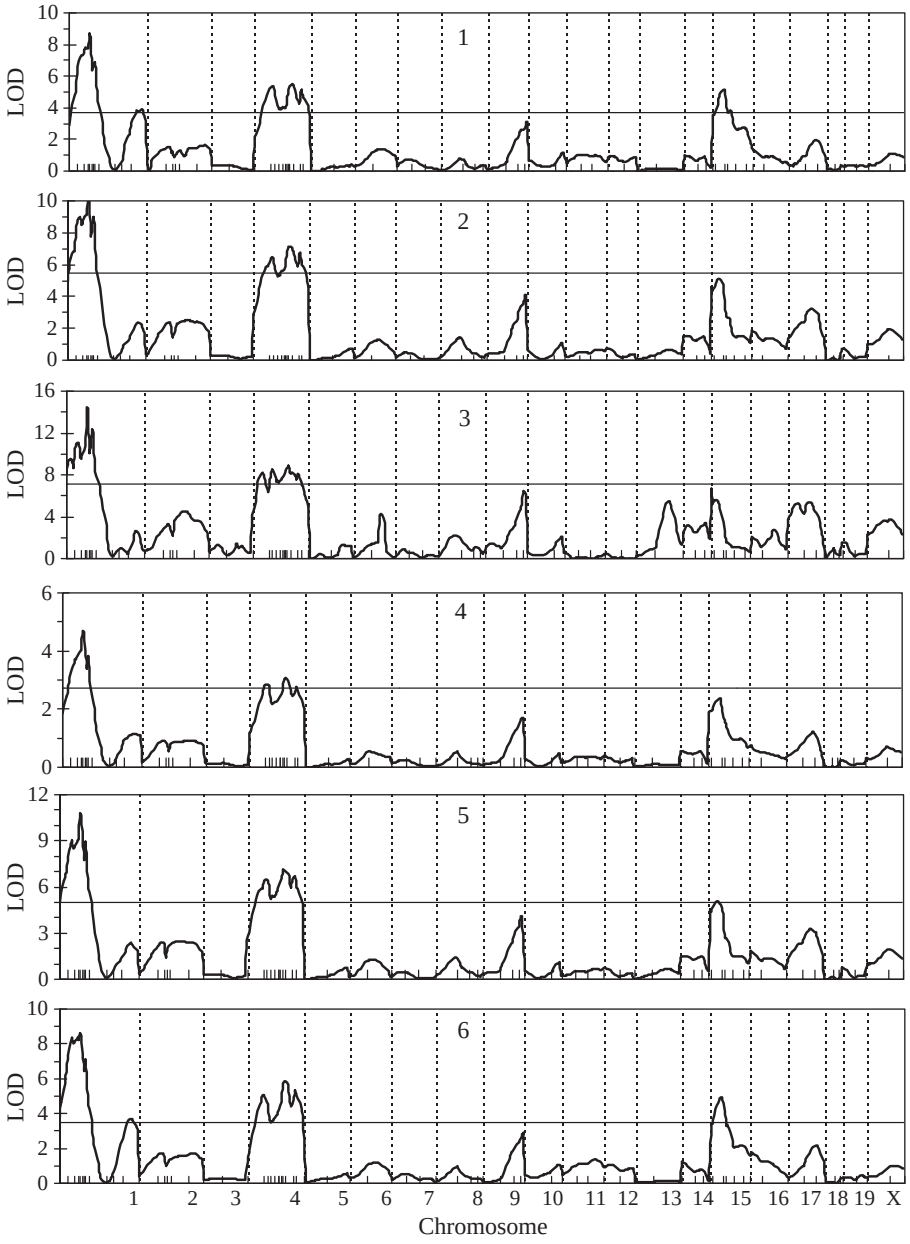

Fig. 1. The profiles of LOD test statistics obtained with the interval mapping based on the six competing survival distributions for HALI survival time in mice: 1. Generalized F distribution; 2. Exponential distribution; 3. Weibull distribution; 4. Log-Normal distribution; 5. Gamma distribution and 6. Log-Logistic distribution. In each plot, the horizontal reference line is the empirical critical value. Linkage groups are separated by the vertical dotted lines and marker positions are indicated by the ticks on the horizontal axis.

uncovered all the QTLs detected with the other five survival distributions. It should be noted that log-logistic distribution also found the same four QTLs as detected with the generalized *F* distribution (See Fig. 1), which does not provide practical evidence about over-fitting of the generalized *F* distribution in mapping survival time loci.

5. Discussion

Combining the advantage of the AFT model and the parametric estimation method in survival analysis, we have proposed to map survival time loci using parametric AFT model with baseline distribution. The generalized *F* distribution is a general form of the

Table 5
Parameter estimation of the QTLs detected with generalized *F* distribution for hyperoxic acute lung injury survival time in mice.

QTL No.	Chr. position	Marker interval	LOD	σ	μ	a	d	v_1	v_2
1	1–24.0	D1Mit478-D1Mit214	8.687	0.429	4.795	−0.084	−0.004	10.182	3.751
2	1–86.0	D1Mit34-D1Mit361	3.927	0.318	4.820	0.045	−0.075	5.676	2.205
3	4–43.7	D4Mit146-D4Mit308	5.503	0.441	4.785	0.066	0.016	10.445	3.897
4	15–15.6	D15Mit175-D15Mit5	5.164	0.392	4.779	0.065	0.007	8.857	3.089

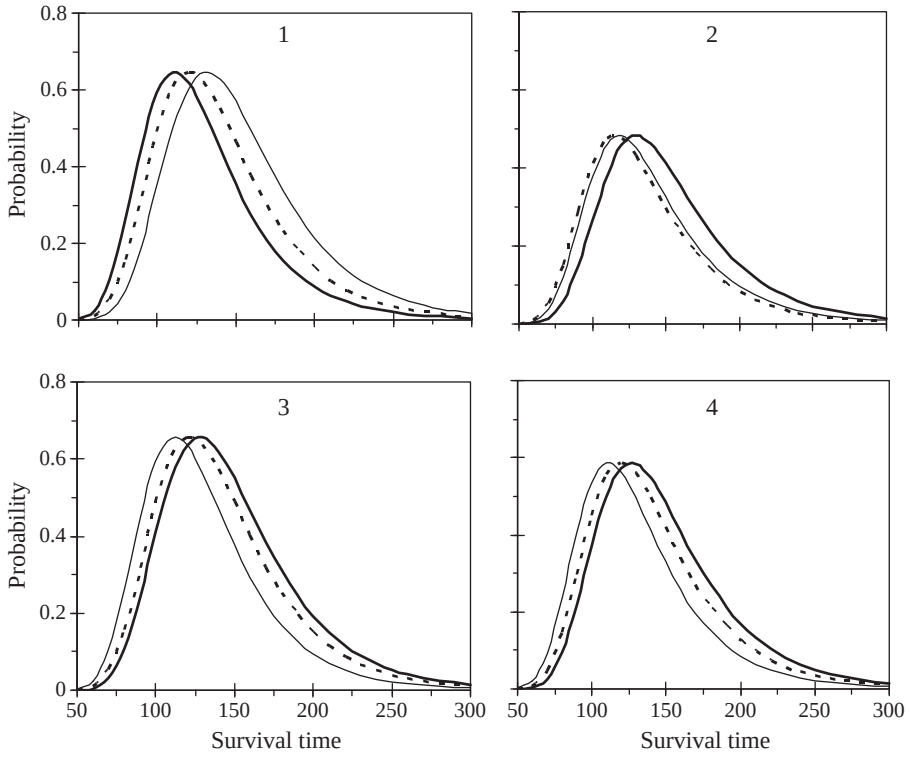

Fig. 2. The survival density curves of three QTL genotypes for 4 detected QTLs (Marked by 1, 2, 3 and 4) drawn according to Formula (4). In each plot, the thick solid, thin solid and dashed lines are for QQ, qq and Qq genotypes, respectively.

commonly used survival distributions. Specifying generalized *F* distribution as the baseline distribution, therefore, we have established a general parametric model for interval mapping of survival traits. The flexibility and performance of the proposed method have been demonstrated through simulation experiments using generalized *F* distribution to fit the data generated from commonly used survival distributions. Real data analysis further shows that generalized *F* distribution cannot only better model HALI survival time than the commonly used survival ones but also can identify all QTLs detected with other competing survival distributions.

Survival traits, in addition to the non-normal distribution, also have a censoring mechanism because of random loss to follow-up, failures from competing causes, or the limited experimental time. Usually, censoring mechanisms are classified into three types: right censored, interval censored and left censored [23]. Let C_{iL} and C_{iR} be the left and right censoring times, respectively, for the *i*th subject. The observation on the trait value of the *i*th subject consists of four components: $t_{iL}=\min\ (T_i,C_{iL})$, $t_{iR}=\max\ (T_i,C_{iR})$, $\Delta_{iL}=I(T_i>C_{iL})$ and $\Delta_{iR}=I(T_i\le C_{iR})$, where $I(\cdot)$ is the indicator function. For mapping QTL using survival data with the censored records, the survival density function (6) in the maximum likelihood estimation should be replaced with the following formula:

$$f(t_i)^{1-I(\cdot)}S(t_i)^{I(\cdot)} = \begin{cases} \frac{1}{\sigma t_i} f_{\varepsilon_i}\left(\frac{logt_i-\mu-z_ia-w_id}{\sigma}\right) & \text{for uncensored} \\ 1-S_{\varepsilon_i}\left(\frac{logt_{iL}-\mu-z_ia-w_id}{\sigma}\right) & \text{for left censored} \\ \left[1-S_{\varepsilon_i}\left(\frac{logt_{iL}-\mu-z_ia-w_id}{\sigma}\right)\right]S_{\varepsilon_i}\left(\frac{logt_{iR}-\mu-z_ia-w_id}{\sigma}\right) & \text{for interval censored} \\ S_{\varepsilon_i}\left(\frac{logt_{iR}-\mu-z_ia-w_id}{\sigma}\right) & \text{for right censored} \end{cases}$$

If the survival distribution is specified as log-normal, then we can treat censored records as missing variables and estimate them with Monte Carlo sampling [33].

Generalized *F* distribution is seldom mentioned in the statistical literature due to its more complicated form than the most commonly used survival distributions and its computational difficulties. Despite this, several advantages are evident when applying the generalized *F* distribution to map survival trait loci, as summarized by Peng et al. [34]. Firstly, it is very flexible and contains other distributions as special cases. Secondly, it can relax the usual stronger distributional assumptions and thirdly it can potentially uncover structure in survival data which otherwise might be missed using other parametric models. For convenience to apply the method, the program implementing model selection for mapping survival trait loci is made in Matlab, which is freely available upon request from the authors.

Acknowledgments

The authors are thankful for the two anonymous reviewers and the editor for helpful comments that have significantly improved the manuscript, and thank for the financial support by the National Natural Science Foundation of China (30972077) to RY.

References

[1] E.S. Lander, D. Botstein, Mapping mendelian factors underlying quantitative traits using RFLP linkage maps, Genetics 121 (1989) 185–199.
[2] Z.B. Zeng, Precision mapping of quantitative trait loci, Genetics 136 (1994) 1457–1468.
[3] C.H. Kao, Z.B. Zeng, R.D. Teasdale, Multiple interval mapping for quantitative trait loci, Genetics 152 (1999) 1203–1216.
[4] S.A. Knott, J.M. Elsen, C.S. Haley, Methods for multiple-marker mapping of quantitative trait loci in half-sib populations, Theor. Appl. Genet. 93 (1996) 71–80.
[5] J.M. Elsen, B. Mangin, B. Goffinet, D. Boichard, P. Le Roy, Alternative models for QTL detection in livestock. I. General introduction, Genet. Sel. Evol. 31 (1999) 213–224.

[6] M.J. Sillanpää, E. Arjas, Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data, Genetics 148 (1998) 1373–1388.
[7] M.J. Sillanpää, E. Arjas, Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data, Genetics 151 (1999) 1605–1619.
[8] J.Y. Cheng, S. Tzeng, Parametric and semiparametric methods for mapping quantitative trait loci, Comput. Statist Data Anal. 53 (2009) 1843–1849.
[9] R.C. Symons, M.J. Daly, J. Fridlyand, T.P. Speed, W.D. Cook, S. Gerondakis, A.W. Harris, S.J. Foote, Multiple genetic loci modify susceptibility to plasmacytoma-related morbidity in E(mu)-v-abl transgenic mice, Proc. Natl Acad. Sci. USA 99 (2002) 11299–11304.
[10] S.R. Lipsitz, J.G. Ibrahim, Estimating equations with incomplete categorical covariates in the Cox model, Biometrics 54 (1998) 1002–1013.
[11] G. Diao, D.Y. Lin, F. Zou, Mapping quantitative trait loci with censored observations, Genetics 168 (2004) 1689–1698.
[12] K.W. Broman, Mapping quantitative trait loci in the case of a spike in the phenotype distribution, Genetics 163 (2003) 1169–1175.
[13] J.P. Fine, F. Zou, B.S. Yandell, Nonparametric estimation of the effects of quantitative trait loci, Biostatistics 5 (2004) 501–513.
[14] L. Kruglyak, E.S. Lander, A nonparametric approach for mapping quantitative trait loci, Genetics 139 (1995) 1421–1428.
[15] G. Diao, D.Y. Lin, Semiparametric methods for mapping quantitative trait loci with censored data, Biometrics 61 (2005) 789–798.
[16] C.R. Moreno, J.M. Elsen, P. Le Roy, V. Ducrocq, Interval mapping methods for detecting QTL affecting survival and time-to-event phenotypes, Genet. Res. 85 (2005) 139–149.
[17] Y. Fang, A note on QTL detecting for censored traits, Genet. Sel. Evol. 38 (2006) 221–229.
[18] M.P. Epstein, X. Lin, M. Boehnke, A tobit variance-component method for linkage analysis of censored trait data, Am. J. Hum. Genet. 72 (2003) 611–620.
[19] V.S. Pankratz, M. de Andrade, T.M. Therneau, Random-effects Cox proportional hazards model: general variance components methods for time-to-event data, Genet. Epidemiol. 28 (2005) 97–109.
[20] Z. Jin, D.Y. Lin, L.J. Wei, Z. Ying, Rank-based inference for the accelerated failure time model, Biometrika 90 (2003) 341–353.
[21] Z.S. Ma, E.J. Bechinski, Accelerated failure time (AFT) modeling for the development and survival of Russian wheat aphid, *Diuraphis noxia* (Mordvilko), Popul. Ecol. 51 (2009) 543–548.
[22] D. Cox, D. Oakes, Analysis of Survival Data, Chapman and Hall, London, 1984.
[23] J.D. Kalbfleisch, R.L. Prentice, The statistical analysis of failure time data, 2nd EdWiley, New York, 2002.
[24] J.Z. Qi, Comparison of Proportional Hazards and Accelerated Failure Time Models, Department of Mathematics and Statistics, University of Saskatchewan, 2009, p. 89.
[25] S.A. Hogg, A. Ciampi, GFREG: a computer program for maximum likelihood regression using the generalized F distribution, Comput. Meth. Programs Biomed. 20 (1985) 201–215.
[26] K.W. Broman, S. Sen, A Guide to QTL Mapping with R/qtl, Springer, 2009.
[27] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Statist. Soc. B. 39 (1977) 1–38.
[28] G.A. Churchill, R.W. Doerge, Empirical threshold values for quantitative trait mapping, Genetics 138 (1994) 963–971.
[29] A.M. Mood, F.A. Graybill, D.C. Boes, Introduction to the Theory of Statistics, 3rd EdMcGraw-Hill, New York, 1974.
[30] D.R. Prows, A.P. Hafertepen, A.V. Winterberg, W.J. Gibbons Jr., C. Liu, T.G. Nick, Genetic analysis of hyperoxic acute lung injury survival in reciprocal intercross mice, Physiol. Genomics 30 (2007) 271–281.
[31] D.R. Prows, A.P. Hafertepen, W.J. Gibbons Jr., A.V. Winterberg, T.G. Nick, A genetic mouse model to investigate hyperoxic acute lung injury survival, Physiol. Genomics 30 (2007) 262–270.
[32] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461–464.
[33] M.J. Sillanpää, F. Hoti, Mapping quantitative trait loci from a single-tail sample of the phenotype distribution including survival data, Genetics 177 (2007) 2361–2377.
[34] Y. Peng, K.B. Dear, J.W. Denham, A generalized F mixture model for cure rate estimation, Stat. Med. 17 (1998) 813–830.